LIDS-P-2272

A GENERAL CLASSIFICATION RULE FOR PROBABILITY MEASURES [1]

Ofer Zeitouni[2]

Department of Electrical Engineering
Technion, Haifa 32000, Israel
zeitouni@ee.technion.ac.il

and

Sanjeev R. Kulkarni[3]

Department of Electrical Engineering
Princeton University,Princeton, NJ 08544
kulkarni@ee.princeton.edu

Submitted: July 29, 1991

Revised: August 12, 1993

ABSTRACT We consider the problem of classifying an unknown probability distribution based on a sequence of random samples drawn according to this distribution. Specifically, if A is a subset of the space of all probability measures $\mathcal{M}_1(\Sigma)$ over some compact Polish space Σ, we want to decide whether or not the unknown distribution belongs to A or its complement. We propose an algorithm which leads a.s. to a correct decision for any A satisfying certain structural assumptions. A refined decision procedure is also presented which, given a countable collection $A_i \subset \mathcal{M}_1(\Sigma)$, $i = 1, 2, \ldots$ each satisfying the structural assumption, will eventually determine a.s. the membership of the distribution in any finite number of the A_i. Applications to density estimation and the problem of order determination of Markov processes are discussed.

Abbreviated Title: Classifying Probability Measures.

Keywords: Hypothesis Testing, Empirical Measure, Large Deviations.

Mathematics Subject Classifications: Primary: 62F03; Secondary: 62G10, 62G20.

[1] This work was supported by the U.S. Army Research Office under grants DAAL03-86-K-0171 (Center for Intelligent Control Systems) and DAAL03-92-G-0320, by the National Science Foundation under grant IRI-9209577, and by the Office of Naval Research under Air Force Contract F19628-90-C-0002.

[2] This work was partially done while the author visited the Laboratory for Information and Decision Systems, MIT

[3] This work was partially done while the author was with the Laboratory for Information and Decision Systems, MIT, and with MIT Lincoln Laboratory, Lexington, MA 02173

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **12 AUG 1993**	2. REPORT TYPE	3. DATES COVERED **00-00-1993 to 00-00-1993**
4. TITLE AND SUBTITLE **A General Classification Rule for Probability Measures**		5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Laboratory for Information and Decision Systems,77 Massachusetts Avenue,Cambridge,MA,02139-4307**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**20**	

1 Introduction

In this paper, we consider the problem of classifying an unknown probability distribution into one of a finite or countable number of classes based on random samples drawn from the unknown distribution. This problem arises in a number of applications involving classification and statistical inference. For example, consider the following problems:

1. Given i.i.d. observations $x_1, x_2, \ldots$ from some unknown distribution P, we wish to decide whether the mean of P is in some particular set (e.g., in some interval or whether the mean is rational, etc.).

2. Given i.i.d. observations $x_1, x_2, \ldots$, we wish to decide whether or not the unknown distribution belongs to a particular parametric class (e.g., to determine if it is Gaussian) or to determine to which of a countable hierarchy of classes the unknown distribution belongs (e.g., to determine class membership based on some smoothness parameter of the density function).

3. We wish to decide whether or not observations $x_1, x_2, \ldots$ are coming from a Markov source, and if so to determine the order of the Markov source.

In these examples, our goal is to decide whether an unknown distribution μ belongs to a set of distributions A or its complement A^c, or more generally to decide to which of a countable collection of sets of distributions $A_1, A_2, \ldots$ the unknown μ belongs. After each new observation x_n we will make a decision as to the class membership of the unknown distribution. Our criterion for success is to require that almost surely only a finite number of mistakes are made. There are two aspects to the "almost sure" criterion. First, as expected, we require that with probability one (with respect to the observations $x_1, x_2, \ldots$) our decision will be correct from some point on. However, depending on the structure of the A_i classification may be difficult for certain distributions μ. Hence, given a measure on the set of distributions we allow failure (i.e., do not require a finite number of mistakes) on a set of distributions of measure zero.

Our work is motivated by the previous work of Cover (1973), Koplowitz (1977), and Kulkarni and Zeitouni (1991). In fact, the previous works just mentioned deal with the specific case in which the unknown distribution is to be classified according to its mean based on i.i.d. observations, as in the example problem 1 above. In this case, a subset of $I\!\!R$ can be identified with the set of distributions A in the natural way (i.e., all distributions whose mean is in a specified set). Cover (1973) considered the case of distributions on $[0,1]$ with $A = \mathcal{Q}_{[0,1]}$, the set of rationals in $[0,1]$, and more generally the case of countable A. He provided a test which, for any measure with mean in A or with mean in $A^c \backslash N$, will make (almost surely) only a finite number of mistakes where N is a set of Lebesgue measure 0. For countable A, Cover also considered the countable hypothesis testing problem of deciding exactly the true mean in the case the true mean belongs to A, and provided a decision rule satisfying a similar success criterion. Koplowitz (1977) showed some properties of sets A which allow for such decision rules and gave some characterizations of the set N. For example, he showed that if $\bar{A}$ (the closure of A) is countable then N is empty, while if $\bar{A}$ is uncountable then N is uncountable. Kulkarni and Zeitouni (1991) extended the results of Cover (1973) by allowing the set A to be uncountable, not necessarily of measure 0, but such that it satisfies a certain structural assumption. Roughly speaking, this structural assumption requires that A be decomposable into a countable union of increasing sets B_m such that a small dilation of B_m increases the Lebesgue measure by only a sufficiently small amount. In a different direction, Dembo and Peres (1991) provide necessary and sufficient conditions for the almost sure discernibility between sets. Their results, when specialized to the set-up discussed above, show that the inclusion of the possibility of some errors on the set of irrationals is necessary in order to ensure discernibility.

The decision rules of [4, 10, 11, 5] are basically as follows. At time n, the smallest m is selected such that the observations are suffiently well-explained by a hypothesis in B_m. If m is not too large, we decide that the unknown distribution belongs to A; otherwise we decide A^c. For the case of countable hypothesis testing, a similar criterion is used. Thus, the B_m can be thougtht of as a decomposition of A into hypotheses of increasing complexity and so the decision rules are reminiscent of Occam's razor or the MDL (Minimum Description Length) principle.

The problem considered in this paper uses a success criterion and decision rules very similar to those in the previous work of [4, 10, 11], but allows much more general types of classification of the unkown distribution. Section 2 treats the case of classification in A versus A^c for distributions on an arbitrary compact complete separable metric space (i.e., a compact Polish space) with i.i.d. observations. The case of classification among a countable number of sets $A_1, A_2, \ldots$ from i.i.d. observations is considered in Section 3. Thus, the results of these two sections cover the example problems 1 and 2 mentioned above. Furthermore, we also consider relaxations of the basic assumption concerning the i.i.d. structure of the observations $x_1, \ldots, x_n$. Namely, results for observations with Markov dependence are presented in Section 4. In particular, we treat example problem 3 on the determination of the order of a Markov chain.

We now give a precise formulation of the problems considered here. Let $x_1, \ldots, x_n$ be i.i.d. samples drawn from some distribution μ (as mentioned, Markov dependence will be considered in Section 4). We assume that x_i takes values in some compact Polish space Σ, which for concreteness should be thought of as $[0,1]^d \subset I\!R^d$. Let $\mathcal{M}_1(\Sigma)$ denote the space of probability measures on Σ. We put on $\mathcal{M}_1(\Sigma)$ the Prohorov metric, denoted $d(\cdot,\cdot)$, whose topology is equivalent to the weak topology.

We consider here the following problems:

P-1) Based on the sequence of observations $(x_1, \ldots, x_n)$, decide whether $\mu \in A$ or $\mu \in A^c$, where A is some given set satisfying certain structural properties (c.f. A-1 below).

P-2) Based on the sequence of observations $(x_1, \ldots, x_n)$, decide whether $\mu \in A_i$ where all $A_i \subset \mathcal{M}_1(\Sigma)$, $i = 1, 2, \ldots$ are sets satisfying structural properties (c.f. A-1 below).

Since $\mathcal{M}_1(\Sigma)$ is a Polish space, there exist on $\mathcal{M}_1(\Sigma)$ many finite measures which we may assume to be normalized to have a total mass 1. Suppose one is given a particular measure, denoted G, on $\mathcal{M}_1(\Sigma)$. In particular, we allow G to charge all open sets in $\mathcal{M}_1(\Sigma)$. G will play the role of the Lebesgue measure in the following structural condition, which is reminiscent of the assumption in Kulkarni and Zeitouni (1991):

A-1) There exists a sequence of open sets $C_m \subset \mathcal{M}_1(\Sigma)$ and closed sets $B_m \subset \mathcal{M}_1(\Sigma)$, and a sequence of positive constants $\epsilon(m)$ such that:

1) $\forall \mu \in A \; \exists m_0(\mu) < \infty \; s.t. \; \forall m > m_0(\mu), \mu \in B_m$.

2) $d(B_m, C_m^c) = \sqrt{2\epsilon(m)} > 0$.

3) $G(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} (C_m^{(\sqrt{2\epsilon(m)})} \backslash A)) = 0$ where $C_m^{(\sqrt{2\epsilon(m)})} = \{\nu \in \mathcal{M}_1(\Sigma) \mid d(\nu, C_m) < \sqrt{2\epsilon(m)}\}$ is the $\sqrt{2\epsilon(m)}$ dilation of C_m.

A-1) is an embellishment of the structural assumption in Kulkarni and Zeitouni (1991), which corresponds to the case where B_m is a monotone sequence and C_m are taken as the $\sqrt{2\epsilon(m)}$ dilation of B_m. The use of **A-1)** 1) and **A-1)** 2) was proposed to us by A. Dembo and Y. Peres, who obtained also various conditions for full discernibility between hypotheses, c.f. Dembo and Peres (1991). We note that as in Kulkarni and Zeitouni (1991), the assumption is immediately satisfied for countable sets A by taking as B_m the union of the first m components of A and noting that, for a finite measure on a metric space, $G(B(x, \delta) \backslash \{x\}) \to_{\delta \to 0} 0$ where $B(x, \delta)$ denotes the open ball of radius δ around x. More generally, A-1) is satisfied for any closed set by taking $B_m = A$ and using for C_m a sequence of open sets which include A whose measure converges to the outer measure of A. Since C_m is open and Σ is compact, it follows that $d(A, C_m^c) > 0$, and A-1) is satisfied. By the same considerations, it also follows that A-1) is satisfied for any countable union of closed sets. Also, note that whenever both $A = \cup_{i=1}^{\infty} A_i$ and $A^c = \cup_{i=1}^{\infty} D_i$ with A_i, D_i closed then, choosing $B_m = \cup_{i=1}^{m} A_i$ and $C_m = \cap_{i=1}^{m} D_i^c$, one sees that A-1) holds (with actually an empty intersection in A-1) 3) for appropriate $\epsilon(m)$). In this situation, the results of this paper correspond to the sufficient part of Dembo and Peres (1991).

2 Classification in A versus A^c

The definition of success of the decision rule will be similar to the one used in Kulkarni and Zeitouni (1991). Namely, a test which makes at each instant n a decision whether $\mu \in A$ or $\mu \in A^c$ based on $x_1, \ldots, x_n$ will be called successful if:

(S.1) $\forall\, \mu \in A$, a.s. ω, $\exists\, T(\omega)$ s.t. $\forall\, n > T(\omega)$, the decision is 'A'.

(S.2) $\exists\, N \subset \mathcal{M}_1(\Sigma)$ s.t.

(S.2.1) $G(N) = 0$

(S.2.2) $\forall\, \mu \in A^c \backslash N$, a.s. ω, $\exists\, T(\omega)$ s.t. $\forall\, n > T(\omega)$, the decision is 'A^c'.

Note that the outcome is unspecified on N. Note also that the definition is asymmetric in the roles played by A, A^c in the sense that errors in A are not allowed at all.

Let $\mu_n = \frac{1}{n} \sum\limits_{i=1}^{n} \delta_{x_i}$. We recall that μ_n satisfies a large deviation principle, i.e.

$$\begin{aligned} -\inf_{\theta \in \bar{A}} H(\theta|\mu) \leq \liminf{}_{n\to\infty} \frac{1}{n} \log\, P(\mu_n \in A) &\leq \limsup{}_{n\to\infty} \tfrac{1}{n} \log\, P(\mu_n \in A) \\ &\leq -\inf_{\theta \in A^0} H(\theta|\mu) \end{aligned} \tag{2.1}$$

where $\bar{A}$ (A^o) denote the closure (interior) of a set $A \subset \mathcal{M}_1(\Sigma)$ in the weak topology, respectively, and

$$H(\theta|\mu) = \begin{cases} \int_\Sigma d\theta \log \dfrac{d\theta}{d\mu} & \text{if } \theta << \mu \\ \infty & \text{otherwise} \end{cases} \tag{2.2}$$

Our decision rule is very similar to that in Kulkarni and Zeitouni (1991). Specifically, we parse the input sequence $x_1, x_2, \ldots$ to form the subsequences

$$X^m \hat{=} (x_{\beta(m-1)+1}, \cdots, x_{\beta(m)}) \tag{2.3}$$

where the choice of the $\beta(m)$ will be given below. The length of the sequence X^m will be denoted by $\alpha(m)$, so that

$$\beta(m) = \sum_{i=1}^{m} \alpha(i), \;\; \beta(0) = 0. \tag{2.4}$$

We will specify the $\beta(m)$ by appropriately selecting the lengths $\alpha(m)$ of the subsequences.

At the end of each subsequence X^m, we form the empirical measure μ_{X^m} based on the data in the subsequence X^m. Namely,

$$\mu_{X^m} = \frac{1}{\alpha(m)} \sum_{i=\beta(m-1)+1}^{\beta(m)} \delta_{x_i} \tag{2.5}$$

Then we make a decision of whether $\mu \in A$ or $\mu \in A^c$ according to whether $\mu_{X^m} \in C_m$ or not. Between parsings, we do not change the decision.

Recall that from the structural assumption A-1), C_m^c is $\sqrt{2\epsilon(m)}$ separated from B^m. Our idea is to choose $\alpha(m)$ sufficiently large such that if the true measure μ is in B_m, then we will have enough data in forming the empirical measure μ_{X^m} to make the probability of an incorrect decision (deciding A^c because $\mu_{X^n} \in C_m^c$) less than $1/m^2$. If $\alpha(m)$ can be chosen in this manner, then for any $\mu \in A$, once $m > m_0(\mu)$ our probability of error at the end parsing interval m is less than $1/m^2$ so that by the Borel-Cantelli lemma we make only finitely many errors.

To show that $\alpha(m)$ can be chosen to satisfy the necessary properties, we will need a strengthened version of the upper bound in Sanov's theorem (2.1). To do that, we use the notion of covering number:

Definition Let $\epsilon > 0$ be given. The covering number of $\mathcal{M}_1(\Sigma)$, denoted $N(\epsilon, \mathcal{M}_1(\Sigma))$, is defined by

$$N(\epsilon, \mathcal{M}_1(\Sigma)) \triangleq \inf \{n | \exists y_1, \ldots, y_n \in \mathcal{M}_1(\Sigma) \quad \text{s.t.} \quad B \subset \cup_{i=1}^{n} B(y_i, \epsilon)\} \tag{2.6}$$

where $B(y, \epsilon)$ denotes a ball of radius ϵ (in the Prohorov metric) around y.

Similarly, for any given ϵ, denote by $N^\Sigma(\epsilon)$ the covering number of Σ, i.e.

$$N^\Sigma(\epsilon) \triangleq \inf \{n | \exists \tilde{y}_1, \ldots, \tilde{y}_n \in \Sigma \quad \text{s.t.} \quad \Sigma \subset \cup_{i=1}^{n} B(\tilde{y}_i, \epsilon)\}. \tag{2.7}$$

where $B(\tilde{y}_i, \epsilon)$ are taken in the metric corresponding to Σ.

We claim now:

Lemma 1

$$N(\epsilon, \mathcal{M}_1(\Sigma)) \leq 2 \left(\frac{e}{\epsilon}\right)^{(N^\Sigma(\epsilon))} \triangleq \bar{N}(\epsilon, \mathcal{M}_1(\Sigma)) \tag{2.8}$$

Proof In order to prove the lemma, we will explicitly construct an ϵ-cover of $\mathcal{M}_1(\Sigma)$ with $\bar{N}(\epsilon, \mathcal{M}_1(\Sigma))$ elements.

Let $\tilde{y}_1, \ldots, \tilde{y}_{N^\Sigma(\epsilon)}$ be the centers of a set of ϵ balls in Σ which create the cover $N^\Sigma(\epsilon)$ in (2.7). Let $\delta_i \triangleq \delta_{\tilde{y}_i}$, i.e. the distribution concentrated at $\tilde{y}_i$, and let

$$\mu_i^j \triangleq j \cdot \left(\frac{\epsilon}{N^\Sigma(\epsilon)}\right) \cdot \delta_i \quad , \quad j = 0, 1, \ldots, \frac{N^\Sigma(\epsilon)}{\epsilon}$$

Define $Y \triangleq \{y \in \mathcal{M}_1(\Sigma) : \exists\, (i_1, j_1) \cdots (i_k, j_k) \text{ s.t. } y = \sum_{\alpha=1}^{k} \mu_{i_\alpha}^{j_\alpha}\}$. Note that Y is a finite set, for it includes at most $\left(\frac{N^\Sigma(\epsilon)}{\epsilon} + 1\right)^{N^\Sigma(\epsilon)}$ members. Also, note that Y is an ϵ-cover of $\mathcal{M}_1(\Sigma)$, i.e. for any

$\mu \in \mathcal{M}_1(\Sigma)$ there exists a $y \in Y$ such that for any open set $C \subset \Sigma$, $\mu(C) \leq y(C^\epsilon) + \epsilon$. To see that, choose as y the following approximation to μ:

Let $i_\alpha = \alpha$, $\alpha = 1, \dots, N^\Sigma(\epsilon)$, and choose $j_\alpha = \left\lfloor \mu\left(B(\tilde{y}_\alpha, \epsilon) \backslash \left(\cup_{k=1}^{\alpha-1} B(\tilde{y}_k, \epsilon)\right)\right)\right\rfloor \frac{N^\Sigma(\epsilon)}{\epsilon}$, where by $\lfloor \times \rfloor$ we mean the closest approximation to $\times$ on the $\frac{N^\Sigma(\epsilon)}{\epsilon}$ j-net from below. Finally, let $j_{N^\Sigma(\epsilon)} \triangleq \frac{N^\Sigma}{\epsilon} - \sum_{\alpha=1}^{N^\Sigma(\epsilon)-1} j_\alpha$. Take now $y = \sum_{\alpha=1}^{N^\Sigma(\epsilon)} \mu_{i_\alpha}^{j_\alpha}$. It follows that y is a probability measure based on a finite number of atoms and, furthermore, $d(y, \mu) < \epsilon$. We need therefore only to estimate the cardinality of the set Y, denoted $|Y|$. Note that $|Y|$ is just the number of vectors $(j_1, \dots, j_{N^\Sigma(\epsilon)})$ such that $\sum_{i=1}^{N^\Sigma(\epsilon)} j_i = 1$ and $j_i \in \{0, \frac{\epsilon}{N(\epsilon)}, \frac{2\epsilon}{N(\epsilon)}, \dots, 1\}$.

It follows that

$$\begin{aligned} |Y| &\leq \left(\tfrac{N^\Sigma(\epsilon)}{\epsilon} + 1\right)^{N^\Sigma(\epsilon)} \int_0^1 \cdots \int_0^{x_3} \int_0^{x_2} dx_1 \dots dx_{N^\Sigma(\epsilon)} \\ &= \left(\tfrac{N^\Sigma(\epsilon)}{\epsilon} + 1\right)^{N^\Sigma(\epsilon)} \cdot \tfrac{1}{N^\Sigma(\epsilon)!} \end{aligned} \tag{2.9}$$

However, by Stirling's formula

$$\log(N^\Sigma(\epsilon)!) \geq N^\Sigma(\epsilon) \log N^\Sigma(\epsilon) - N^\Sigma(\epsilon) \tag{2.10}$$

Substituting (2.10) into (2.9), one has

$$|Y| \leq \left(\frac{N^\Sigma(\epsilon)}{\epsilon} + 1\right)^{N^\Sigma(\epsilon)} e^{N^\Sigma(\epsilon)} \cdot \frac{1}{(N^\Sigma(\epsilon))^{N^\Sigma(\epsilon)}} \tag{2.11}$$

which implies that

$$\begin{aligned} N(\epsilon, \mathcal{M}_1(\Sigma)) &\leq \left(\tfrac{1}{\epsilon}\left(1 + \tfrac{\epsilon}{N^\Sigma(\epsilon)}\right)\right)^{N^\Sigma(\epsilon)} e^{N^\Sigma(\epsilon)} \\ &= \left(\tfrac{e}{\epsilon}\right)^{N^\Sigma(\epsilon)} \left(1 + \tfrac{\epsilon}{N^\Sigma(\epsilon)}\right)^{N^\Sigma(\epsilon)} \leq 2\left(\tfrac{e}{\epsilon}\right)^{N^\Sigma(\epsilon)} = \bar{N}(\epsilon, \mathcal{M}_1(\Sigma)) \end{aligned}$$

□

For completeness, we show in the Appendix a complementary lower bound on the covering number which exhibits a behavior similar to $\bar{N}$. Thus, the upper bound $\bar{N}$ cannot be much improved.

The existence of the bound $\bar{N}$ permits us to mimic the computation in Kulkarni and Zeitouni (1991) for the case in hand. Indeed, a crucial step needed is bounding the probability of complements of balls, for all n, uniformly over all measures, as follows:

Theorem 1

$$P(\mu_n \in B(\mu, \delta)^c) \leq \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) e^{-n\left(\frac{\delta}{4}\right)^2}$$

Proof The proof follows the standard Chebycheff bound technique, without taking n limits as in the large deviation framework. Indeed,

$$P(\mu_n \in B(\mu,\delta)^c) \leq \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \cdot \sup_{y \in \mathcal{M}_1(\Sigma), d(y,\mu) \geq 3\delta/4} P(\mu_n \in B(y, \frac{\delta}{4}))$$

Therefore, by the Chebycheff bound, denoting by P_n the law of the random variable μ_n and by $C_b(\Sigma)$ the space of continuous functions on Σ, it follows that for any $\theta \in C_b(\Sigma)$,

$$\begin{aligned}
P(\mu_n \in B(\mu,\delta)^c) &\leq \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \cdot \sup_{y \in \mathcal{M}_1(\Sigma), d(y,\mu) \geq 3\delta/4} \int_{B(y,\frac{\delta}{4})} e^{n<\theta,\nu>} e^{-n<\theta,\nu>} dP_n(\nu) \\
&\leq \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \cdot \sup_{y:\, d(y,\mu) \geq 3\delta/4} \exp\left(-n \sup_{\theta \in C_b(\Sigma)} \inf_{\nu \in B(y,\frac{\delta}{4})} (<\theta,\nu> - \frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>}))\right) \\
&= \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \cdot \exp\left(-n \inf_{\nu \in B(y,\frac{\delta}{4}), d(y,\mu) \geq 3\delta/4} \sup_{\theta \in C_b(\Sigma)} (<\theta,\nu> - \frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>}))\right) \\
&= \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \cdot \exp\left(-n \inf_{\nu \in B(y,\frac{\delta}{4}), d(y,\mu) \geq 3\delta/4} H(\nu|\mu)\right) \\
&\leq \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \cdot \exp\left(-n \inf_{\nu \in B(\mu,\frac{\delta}{2})^c} H(\nu|\mu)\right) \\
&\leq \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) \cdot e^{-n\,(\frac{\delta}{4})^2} \qquad (2.12)
\end{aligned}$$

where $<\theta,\nu>= \int \theta(x)\nu(dx)$, the first equality in (2.12) follows from the min-max theorem for convex compact sets (c.f. Theorem 4.2 of Sion (1958)), the second equality follows by Lemma 3.2.13 of Deuschel and Stroock (1989), and the last inequality from the fact that (Deuschel and Stroock (1989), Exercise 3.2.24) for any $\theta \in B(\mu,\delta/2)^c$,

$$\frac{\delta}{2} \leq d(\theta,\mu) \leq \|\theta - \mu\|_{var} \leq 2H^{1/2}(\theta|\mu)$$

□

Corollary 1 Let $B_m \subset \mathcal{M}_1(\Sigma)$ be a measurable set such that $\mu \in B_m$. Let B_m^δ denote an open set such that $d(B_m,(B_m^\delta)^c) \geq \delta$. Then

$$P\left(\mu_n \in (B_m^\delta)^c\right) \leq \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma)\right) e^{-n(\frac{\delta}{4})^2} \qquad (2.13)$$

We return now to the proposed classification algorithm. Motivated by Corollary 1, define

$$\alpha(m) = \frac{8}{\epsilon(m)}\left[2\log m + \log 2 + N^{\Sigma}\left(\sqrt{\frac{\epsilon(m)}{8}}\right)\left(1 - \log\sqrt{\frac{\epsilon(m)}{8}}\right)\right] \qquad (2.14)$$

and let $\beta(m)$ be as defined previously by (2.4).

Note that with this choice of $\alpha(m)$, using Corollary 1 with $\delta = \sqrt{2\epsilon(m)}$, and the expression for $\bar{N}(\delta/4, \mathcal{M}_1(\Sigma))$ from Lemma 2.8, we have that for all $\mu \in A$ and $m > m_0(\mu)$,

$$P(\mu_{\alpha(m)} \in C_m^c) \leq \frac{1}{m^2} \tag{2.15}$$

as we wanted.

For convenience we summarize the decision rule again here.

Decision Rule For any input sequence $x_1, x_2, \ldots$, form the subsequences

$$X^m \hat{=} (x_{\beta(m-1)+1}, \cdots, x_{\beta(m)}).$$

Let μ_{X^m} denote the empirical measure of the sequence X^m. At the end of each parsing, decide $\mu \in A$ if $\mu_{X^m} \in C_m$ and decide $\mu \in A^c$ otherwise. Between parsings, don't change the decision.

We now claim:

Theorem 2 The decision rule defined by the parsing $\beta(m)$ as above is successful.

Proof The proof is essentially identical to the proof of Theorem 1 in Kulkarni and Zeitouni (1991).

a) If $\mu \in A$, then by assumption **A-1)**1) there exists $m_0(\mu)$ such that $\mu \in B_m$ for all $m > m_0(\mu)$. Note that the event of making an error infinitely often is equivalent to the event of making an error **at the parsing intervals** infinitely often. However, by our choice of $\alpha(m)$

$$\sum_{m=1}^{\infty} \text{Prob}\{\text{error after } m\text{−th parsing}\} \leq m_0(\mu) + \sum_{m=m_0(\mu)+1}^{\infty} \frac{1}{m^2} < \infty$$

Therefore, using the Borel-Cantelli lemma, we have that our decision rule satisfies part (S.1) of the definition of a successful decision rule.

b) Let

$$N = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} C_m^{(\sqrt{2\epsilon(m)})} \backslash A \tag{2.16}$$

By assumption **A-1)**3), $G(N) = 0$. Now, if $\mu \in A^c \backslash N$, we may repeat the arguments of part a) in the following way: For an $m_0(\mu)$ large enough, $\mu \in (C_m^{(\sqrt{2\epsilon(m)})})^c$ for all $m > m_0(\mu)$. Therefore, we have $d(\mu, C_m) \geq \sqrt{2\epsilon(m)}$ for all $m > m_0(\mu)$. Then using Corollary 1 with $\delta = \sqrt{2\epsilon(m)}$, the expression for $\bar{N}(\delta/4, \mathcal{M}_1(\Sigma)$ from Lemma 2.8, and the choice of $\alpha(m)$ we have that for $m > m_0(\mu)$

$$\text{Prob}\{\text{error after } m\text{−th parsing}\} = P(\mu_{X^m} \in C_m) \leq \frac{1}{m^2} \tag{2.17}$$

Hence, as in part a), the result follows by a smiple application of the Borel-Cantelli lemma.

□

3 Classification Among a Countable Number of Sets

In this section, we refine the decision rule to allow for classification among a countable number of sets. Specifically, if $A_1, A_2, \ldots$ are a countable number of subsets of $\mathcal{M}^1(\Sigma)$ we are interested in deciding to which of the A_i the unknown measure μ belongs. The only assumption we make on the A_i is that each A_i satisfies the structural assumption (A-1). The A_i are not required to be either disjoint or nested, although these special cases are most commonly of interest in applications. In general, after a finite number of observations one cannot expect to determine the membership status of μ in all of the A_i. However, we will show that for all μ except in a set of G-measure zero in $\mathcal{M}^1(\Sigma)$ there is a decision procedure that a.s. will eventually determine the membership of μ in any finite subset of the A_i. In the special cases of disjoint or nested A_i, the membership status of μ in any of the countable A_i is completely determined by membership in some finite subset. Hence, in these cases, except for μ in a set of G-measure zero the membership of μ in all the A_i will a.s. be eventually determined.

We modify our previous decision rule as follows. The observations $x_1, x_2, \ldots$ will still be parsed into increasingly larger blocks in a manner to be defined below. However, now, at the end of the m-th block, we will make a decision as to the membership of μ in the first m of the A_i. The decisions of whether μ belongs to $A_1, \ldots, A_m$ are made separately for each A_i using a procedure similar to that of the previous section.

Specifically, for each A_i let $B_{i,m}$ be a sequence of closed sets, $C_{i,m}$ a sequence of open sets and $\epsilon_i(m) \to_{m\to\infty} 0$ a positive sequence satisfying the requirements of the structural assumption (A-1). From the same considerations that led to (2.15), for

$$\alpha_i(m) = \frac{8}{\epsilon_i(m)}\left[2\log m + \log 2 + N^{\Sigma}\left(\sqrt{\epsilon_i(m)/8}\right)\left(1 - \log\sqrt{\epsilon_i(m)/8}\right)\right] \tag{3.18}$$

we have, for $\mu \in A_i$,

$$P_\mu(\mu_{\alpha_i(m)} \in C_{i,m}^c) \le \frac{1}{m^2} \tag{3.19}$$

As before, the observation sequence $x_1, x_2, \ldots$ will be parsed into non-overlapping blocks

$$X^m = (x_{\beta(m-1)+1}, \ldots, x_{\beta(m)}) \tag{3.20}$$

where the $\beta(m)$ are defined below. At the end of the m-th block, a decision will be made about the membership of μ in $A_1, \ldots, A_m$. This decision will be made separately for each $i = 1, \ldots, m$ using the observation sequence X^m exactly as before. That is, at the end of the parsing sequence X^m, for $i = 1, \ldots, m$ decide that $\mu \in A_i$ according to whether or not $\mu_{X^m} \in C_{i,m}$, and don't change the decision except at the end of a parsing sequence. We define the parsing sequence $\beta(m)$ by $\beta(0) = 0$ and $\beta(m) - \beta(m-1) =$

$\max_{1\leq i\leq m} \alpha_i(m)$ or equivalently

$$\beta(m) = \sum_{k=1}^{m} \max_{1\leq i\leq k} \alpha_i(k), \qquad \beta(0) = 0 \tag{3.21}$$

For this decision rule we have the following theorem.

Theorem 3 Let $A_i \subset \mathcal{M}^1(\Sigma)$ for $i = 1, 2, \ldots$ satisfy the structural assumption (A-1). There is a set $N \subset M_1(\Sigma)$ of G-measure zero such that for every $\mu \in \mathcal{M}^1(\Sigma) \setminus N$ and every $k < \infty$ the decision rule will make (a.s.) only a finite number of mistakes in deciding the membership of μ in $A_1, \ldots, A_k$. That is, given any $\mu \in \mathcal{M}^1(\Sigma) \setminus N$, for a.e. ω there exists $m(\omega) = m(\omega, \mu, k)$ such that for all $m > m(\omega)$ the algorithm makes a correct decision as to whether $\mu \in A_i$ or $\mu \in A_i^c$ for $i = 1, \ldots, k$.

Proof Let

$$N_i = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} C_{i,m}^{(\sqrt{2\epsilon(m)})} \setminus A_i \tag{3.22}$$

and let

$$N = \bigcup_{i=1}^{\infty} N_i \tag{3.23}$$

Then from the assumption (A-1) it follows that the G-measure of each N_i is zero, and so the G-measure of N is also zero.

Now, let $\mu \in \mathcal{M}^1(\Sigma) \setminus N$, and let $k < \infty$. For each $i = 1, \ldots, k$, there exists $m_i(\mu) < \infty$ such that if $\mu \in A_i$ then $\mu \in B_{i,m}$ for all $m > m_i(\mu)$, while if $\mu \in A_i^c$ then $\mu \in (C_{i,m}^{(\sqrt{2\epsilon(m)})})^c$ for all $m > m_i(\mu)$ (since $\mu \notin N_i$). Recall that at the end of the parsing sequence X^m, the algorithm decides $\mu \in A_i$ iff $\mu_{X^m} \in C_{i,m}$, so that if $\mu \in A_i$ then an error is made about membership in A_i iff $\mu_{X^m} \notin C_{i,m}$ while if $\mu \notin A_i$ an error is made iff $\mu_{X^m} \in C_{i,m}$. If $\mu \in A_i$ then using Corollary 1 and the fact that $d(B_{i,m}, C_{i,m}^c) \geq \sqrt{2\epsilon_i(m)}$, we have that the probability of making an incorrect decision is less than $1/m^2$ for $m > m_i(\mu)$. On the other hand, if $\mu \in A_i^c$ then since $d(C_{i,m}, (C_{i,m}^{(\sqrt{2\epsilon(m)})})^c) \geq \sqrt{2\epsilon_i(m)}$ we also have probability of error less than $1/m^2$ for $m > m_i(\mu)$ (again using Corollary 1 and the expression for $\alpha(m)$). Hence, for $m > m_0(\mu) = \max(m_{(}\mu), \ldots, m_k(\mu))$ the probability of making an error about the membership of μ in *any* of $A_i, \ldots, A_k$ is less than k/m^2. Then

$$\sum_{m=1}^{\infty} \text{Prob}\{\text{error in any } A_i \text{ on } m\text{-th parsing}\} \leq m_0 + k \sum_{m=m_0+1}^{\infty} \frac{1}{m^2} < \infty$$

so that the theorem follows by the Borel-Cantelli Lemma.

□

Note that if one also wants to make a correct decision after some finite time whether or not μ is in *any* of the A_i for $i = 1, 2, \ldots$ then the decision procedure can be easily modified to handle this. Specifically, it is easy to show that sets satisfying the structural assumption are closed under countable union. Hence, one could include in the hypothesis testing the set $A_0 = \cup_{i=1}^{\infty} A_i$, so that after some finite time a correct decision would be made about the membership of $\mu \in A_0$.

Also, it is worthwhile to note that if the A_i have more structure then some improvements can be made. For example, if the membership status of μ in A_i for $i = 1, 2, \ldots$ is determined by its membership status in some finite number of the A_i then a correct decision regarding the membership of μ in all of the A_i can be guaranteed (a.s.) after some finite time (depending on μ). This is the case for disjoint or nested A_i, which may be of particular interest in some applications. For these cases, by letting $A_0 = \cup_{i=1}^{\infty} A_i$ and running the decision rule on $A_0, A_1, A_2, \ldots$ as mentioned above, we have the following corollary of Theorem 3.

Corollary 2 Let $A_i \subset \mathcal{M}^1(\Sigma)$ for $i = 1, 2, \ldots$ satisfy the structural assumption (A-1) and suppose the A_i are either disjoint or nested. There is a set $N \subset M_1(\Sigma)$ of G-measure zero such that for every $\mu \in \mathcal{M}^1(\Sigma) \backslash N$ the decision rule will make (a.s.) only a finite number of mistakes in deciding the membership of μ in all of the A_i. That is, given any $\mu \in \mathcal{M}^1(\Sigma) \setminus N$, for a.e. ω there exists $m(\omega) = m(\omega, \mu)$ such that for all $m > m(\omega)$ the algorithm makes a correct decision as to whether $\mu \in A_i$ for all $i = 1, 2, \ldots$.

It is worthwhile to note that the results of this section may be used also in the case that Σ is locally compact but not compact. In that case, one may first intersect the A_i with compact sets K_m which sequentially approximate Σ and then use $m(n) \to \infty$. We do not consider this issue here.

We conclude this section with an example taken from the problem of density estimation. Let $\Sigma = [0, 1]$ and assume that $x_1, \ldots, x_n$ are i.i.d. and drawn from a distribution with law μ_θ, $\theta \in \Theta$. When some structure is given on the set $\mathcal{F} = \bigcup_{\theta \in \Theta} \mu_\theta$, there exists a large body of literature which enables one to obtain estimates of the error after n observations (e.g., see Ibragimov and Has'minskii (1981)). All these results assume an a-priori structure, e.g. a bound on the L^2 norm of the density $f_\theta = \frac{d\mu_\theta}{dx}$. If such information is not given a-priori, it may be helpful to design a test to check for this information and thus to be able to estimate eventually whether the distribution belongs to a nice set and if so to apply the error estimates alluded to above. The application of such an idea to density estimation was suggested by Cover (1972).

As a specific example, let

$$A_i = \{\mu \in \mathcal{M}_1(\Sigma) : \int_0^1 (\frac{d\mu(x)}{dx})^2 \leq i\}.$$

Note that the sets A_i are closed w.r.t. the Prohorov metric and therefore they satisfy the structural assumption A-1). Moreover, they are nested and thus Corollary 2 may be applied to yield a decision rule which will asymptotically decide correctly on the appropriate class of densities.

A somewhat different application to density estimation arises when the A_i consist of single points (i.e., each A_i contains a single probability measure). The special case in which A_i consists of the i-th computable density is related to a model considered by Barron (1985) and Barron and Cover (1991). For an estimation procedure based on the Minimum Description Length (MDL) principle, they showed strong consistency results when the true density is a computable one. Since, there are a countable number of computable densities and the structural assumption A-1) is satisfied for any singleton, a strong consistency result for computable densities follows immediately from our results.

4 Applications to Order Determination of Markov Processes

In this section, we extend the model of the observations to allow for a Markov dependence in the observation. The problem we wish to consider is the order selection problem: given observations from an (unknown) Markov chain, one wishes to estimate the order of the chain in order to best fit a Markov model to the data.

Specifically, let Σ be a compact Polish space as before, but assume that the observations $x_1, \dots, x_n$ are the outcome of a Markov chain of order j, i.e.

$$\mathrm{Prob}(x_k \in A | x_{k-1}, x_{k-2}, \dots, x_1) = \pi^j(x_k \in A | x_{k-1}, x_{k-2}, \dots, x_{k-j})$$

where A is a Borel measurable subset of Σ and $k > j$. In order to avoid technicalities, we assume that all Markov chains involved are ergodic, and therefore there exists a unique stationary measure $P_{\pi^j} \in \mathcal{M}_1(\Sigma^j)$ such that for any measurable set A in Σ^j,

$$P_{\pi^j}(A) = \int_{\{x_{2j}, x_{2j-1}, \dots, x_{j+1}\} \in A} d\pi^j(x_{2j} | x_{2j-1}, \dots, x_j) \cdots d\pi^j(x_{j+1} | x_j, \dots, x_1) dP_{\pi^j}(x_j, \dots, x_1) \qquad (4.24)$$

We assume that j is unknown, and our task is to decide (correctly) on the order j.

This problem has already been considered in the literature. Hannan and Quinn (1979) and later Hannan (1980) considered the case of autoregressive and ARMA models, and proved, under some assumptions, the consistency of an estimator based on the Akaike criterion. For a related work, see Shibata (1980). In all the above, an effort is made also to prove asymptotic optimality of the proposed estimators. In the discrete alphabet (finite Σ) set-up, Merhav et al. (1989) proposed an estimator based on relative entropy, related it to the Lempel-Ziv compression algorithm, and proved its asymptotic optimality in the sense of large deviations. However, their approach does not guarantee in general a zero probability of error and may result in biased estimates.

In this section, we depart from the above by, on the one hand, relaxing the requirement for "asymptotic optimality" and, on the other hand, considering the general setup of Markov chains. We show how a strongly consistent decision rule may be constructed based on the general paradigm of this paper. Towards this end, we need to extend the basic estimates of Section 2 to the Markov case, as follows.

Let $\Omega = \Sigma^Z$, define x_i to be the coordinate map $x_i(\omega) = \omega_i$, and let the shift operator be defined by $x_i(T\omega) = x_{i+1}(\omega)$. Define the k-th order empirical measure on $\mathcal{M}_1(\Sigma^k)$ by

$$\mu_n^k = \mu_n^k(\omega) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_1(T^i\omega), x_2(T^i\omega), \ldots, x_k(T^i\omega)}$$

As before, we endow $\mathcal{M}_1(\Sigma^k)$ with the Prohorov topology, and recall that a large deviations upper bound holds for the empirical measure μ_n^{j+1}, viz. for any set $A \subset \mathcal{M}_1(\Sigma^{j+1})$ a large deviations statement of the form (2.1) holds, with the relative entropy $H(\nu|\mu)$ being replaced by

$$H_j(\theta|\mu) = \begin{cases} \int_{\Sigma^j} d\theta(y_1, \ldots, y_{j+1}) \log \dfrac{d\theta(y_{j+1}|y_j, \ldots, y_1)}{d\mu(y_{j+1}|y_j, \ldots, y_1)} & \text{if } \theta(\cdot|y_j, \ldots, y_1) << \mu(\cdot|y_j, \ldots, y_1) \\ \infty & \text{otherwise} \end{cases} \tag{4.25}$$

For any measure $\mu(x_1, \ldots, x_k) \in \mathcal{M}_1(\Sigma^k)$, denote by μ_i the marginal defined by

$$\mu_i(\{x_1, \ldots, x_i\} \in A) \triangleq \mu(\{x_1, \ldots, x_i\} \in A, \{x_{i+1}, \ldots, x_k\} \in \Sigma^{k-i})$$

and by $\mu_{i|i-1,\ldots,i-t}$ the regular conditional probability $\mu(x_i|x_{i-1}, \ldots, x_{i-t})$. With a slight abuse of notations, we continue to use μ_i for the marginal of a measure $\mu \in \mathcal{M}_1(\Sigma^Z)$. Define the measure $\bar{\mu} = \mu_{i-k} \otimes \mu_{i-(k-1)|i-k,\ldots,1} \otimes \cdots \otimes \mu_{i|i-1,\ldots,1} \in \mathcal{M}_1(\Sigma^i)$ as the measure which, for any measurable set $A \subset \Sigma^i$,

$$\bar{\mu}(A) = \int_A d\mu_{i-k}(x_1, \ldots, x_{i-k}) d\mu_{i-(k-1)|i-k,\ldots,1}(x_{i-(k-1)}|x_{i-k}, \ldots, x_1) \cdots d\mu_{i|i-1,\ldots,1}(x_i|x_{i-1}, \ldots, x_1) \tag{4.26}$$

Let π^j be a given j-th order Markov kernel, P_{π^j} its corresponding stationary measure, and denote by Pr^{π^j} the stationary measure on Ω generated by this kernel. Assume that the empirical measures μ_n^{j+k}, $k = 2, 3, \ldots$ are formed from a Markov sequence generated by this kernel. In order to compute explicitly the sequence of decision rules as in the i.i.d. case, we need to derive the analog of Theorem 1 given below.

Theorem 4

$$\begin{aligned} Pr^{\pi^j}\left[\mu_n^{j+k} \notin B\left((\mu_n^{j+2})_j \otimes \pi^j \otimes \cdots \otimes \pi^j, \delta\right)\right] &\leq \\ \bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma^{j+k})\right) e^{-n(\frac{\delta}{16})^2} &+ \cdots + \bar{N}\left(\frac{\delta}{2^{2k+2}}, \mathcal{M}_1(\Sigma^{j+1})\right) e^{-n(\frac{\delta}{2^{2k+4}})^2} \\ &\leq k\bar{N}\left(\frac{\delta}{2^{2k+2}}, \mathcal{M}_1(\Sigma^{j+k})\right) e^{-n(\frac{\delta}{2^{2k+4}})^2} \end{aligned} \tag{4.27}$$

Proof We prove the Theorem first for the case $k = 2$. The general case follows by induction. For any $\nu \in \mathcal{M}_1(\Sigma^{j+2})$, let

$$\mathcal{A}(\nu, \delta) = \{\mu \in \mathcal{M}_1(\Sigma^{j+2}) : d(\mu, \nu_j \otimes \pi^j \otimes \pi^j) < \delta\}$$

$$\mathcal{C}(\nu, \delta) = \{\mu \in \mathcal{M}_1(\Sigma^{j+2}) : d(\mu_{j+1}, \nu_j \otimes \pi^j) < \delta/4\}.$$

It follows that

$$\begin{aligned} Pr^{\pi^j}\left[\mu_n^{j+2} \notin B\left((\mu_n^{j+2})_j \otimes \pi^j \otimes \pi^j, \delta)\right)\right] &\leq Pr^{\pi^j}(\mu_n^{j+2} \notin \mathcal{A}(\mu_n^{j+2}, \delta), \mu_n^{j+2} \in \mathcal{C}(\mu_n^{j+2}, \delta)) \\ &\quad + Pr^{\pi^j}(\mu_n^{j+2} \notin \mathcal{C}(\mu_n^{j+2}, \delta)) \\ &\triangleq P_1 + P_2 \end{aligned} \tag{4.28}$$

By repeating the argument in (2.12),

$$\frac{P_1}{\bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma^{j+2})\right)} \leq \sup_{y \in \mathcal{M}_1(\Sigma^{j+2})} P\left[\mu_n^{j+2} \in B(y, \frac{\delta}{4}), \mu_n^{j+2} \in \mathcal{C}(\mu_n^{j+2}, \delta), y \notin \mathcal{A}(\mu_n^{j+2}, 3\delta/4)\right] \tag{4.29}$$

Therefore, by the Chebycheff bound, denoting by P_n the law of the random variable μ_n^{j+2}, it follows that for any $\theta \in C_b(\Sigma^{j+2})$,

$$\begin{aligned} \frac{P_1}{\bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma^{j+2})\right)} &\leq \sup_{y \in \mathcal{M}_1(\Sigma^{j+2})} \int_{B(y,\frac{\delta}{4})} e^{n<\theta,\nu>} e^{-n<\theta,\nu>} 1_{\mu_n^{j+2} \in \mathcal{C}(\mu_n^{j+2},\delta), y \notin \mathcal{A}(\mu_n^{j+2}, 3\delta/4))} dP_n(\nu) \\ &\leq \sup_{y \in \mathcal{M}_1(\Sigma^{j+2})} \exp\left[-n \sup_{\theta \in C_b(\Sigma^{j+2})} \inf_{\substack{\nu \in B(y,\frac{\delta}{4}) \bigcap \mathcal{C}(\nu,\delta) \\ y \notin \mathcal{A}(\nu, 3\delta/4)}} \left(< \theta, \nu > -\frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>})\right)\right] \\ &= \exp\left[-n \inf_{\substack{\nu \in B(y,\frac{\delta}{4}) \bigcap \mathcal{C}(\nu,\delta) \\ y \notin \mathcal{A}(\nu, 3\delta/4)}} \sup_{\theta \in C_b(\Sigma^{j+2})} \left(< \theta, \nu > -\frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>})\right)\right] \end{aligned} \tag{4.30}$$

However,

$$\begin{aligned} &\sup_{\theta \in C_b(\Sigma^{j+2})} \left(< \theta, \nu > -\frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>})\right) \\ = &\sup_{\theta \in C_b(\Sigma^{j+2})} \left(< \theta, \nu > -\frac{1}{n} \log E_{P_n}(e^{\theta(x_1,\ldots,x_{j+2}) + \cdots + \theta(x_{n-j-2},\ldots,x_n)})\right) \\ = &\sup_{\theta \in \mathcal{B}(\Sigma^{j+2})} \left(< \theta, \nu > -\frac{1}{n} \log E_{P_n}(e^{\theta(x_1,\ldots,x_{j+2}) + \cdots + \theta(x_{n-j-2},\ldots,x_n)})\right) \end{aligned} \tag{4.31}$$

where $\mathcal{B}(\Sigma^{j+2})$ denotes the space of bounded measurable functions on Σ^{j+2} and the last equality follows from dominated convergence. We assume now that ν is absolutely continuous w.r.t $\nu_{j+1} \otimes \pi^j$ and that the resulting Radon-Nikodym derivative is uniformly bounded from above and below (these assumptions may be relaxed exactly as in Deuschel and Stroock (1989), pg. 69). In this case, we may take θ in (4.32) as this Radon-Nikodym derivative, i.e. $\theta(x_1, \ldots, x_{j+2}) = \log \frac{d\nu}{d\nu_{j+1} \otimes \pi^j}$ to obtain:

$$\sup_{\theta \in \mathcal{B}(\Sigma^{j+2})} \left(< \theta, \nu > -\frac{1}{n} \log E(e^{\theta(x_1,\ldots,x_{j+2}) + \cdots + \theta(x_{n-j-1},\ldots,x_n)})\right) \geq H(\nu | \nu_{j+1} \otimes \pi^j) \tag{4.32}$$

Substituting (4.32) in (4.30) and recalling the inequality

$$2H^{1/2}(\nu|\mu) \geq d(\nu, \mu)$$

one obtains

$$\begin{aligned}
\frac{P_1}{\bar{N}\left(\frac{\delta}{4}, \mathcal{M}_1(\Sigma^{j+2})\right)} &\leq \exp\left(-n \inf_{\substack{\nu \in B(y,\frac{\delta}{4}) \bigcap \mathcal{C}(\nu,\delta) \\ y \notin \mathcal{A}(\nu,3\delta/4)}} H(\nu|\nu_{j+1} \otimes \pi^j)\right) \\
&\leq \exp\left(-n \inf_{\substack{\nu \in B(y,\frac{\delta}{4}) \bigcap \mathcal{C}(\nu,\delta) \\ y \notin \mathcal{A}(\nu,3\delta/4)}} d(\nu, \nu_{j+1} \otimes \pi^j)^2/4\right) \\
&\leq \exp\left(-\frac{n}{4} \inf_{\substack{\nu \in B(y,\frac{\delta}{4}) \bigcap \mathcal{C}(\nu,\delta) \\ y \notin \mathcal{A}(\nu,3\delta/4)}} \left(d(\nu, \nu_j \otimes \pi^j \otimes \pi^j) - d(\nu_{j+1} \otimes \pi^j, \nu_j \otimes \pi^j \otimes \pi^j)\right)_+^2\right) \\
&\leq \exp\left(-n\ (\delta/16)^2\right)
\end{aligned} \tag{4.33}$$

Similarly,

$$\frac{P_2}{\bar{N}\left(\frac{\delta}{16}, \mathcal{M}_1(\Sigma^{j+1})\right)} \leq \exp\left(-n\ (\delta/64)^2\right) \tag{4.34}$$

Substituting (4.33), and (4.34) in (4.28) yields the Theorem for $k = 2$. The general case is similar and follows by induction.

□

We are now ready to return to the order determination problem described in the beginning of this section. Since the set up here differs slightly from the one described in the previous section, we repeat here the main definitions.

Let $A_i \subset \mathcal{M}_1(\Omega)$, $i = 0, 1, \ldots$, be the set of stationary measures generated by Markov chains of order i (with $i = 0$ denoting the i.i.d. case), i.e. for $i = 0, 1, 2, \ldots$,

$$\mu \in A_i \Longleftrightarrow (\mu)_{i+k} = (\mu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i \text{ for some Markov kernel } \pi^i \text{ and for all } k = 1, 2, \ldots.$$

Note that the sets A_i are closed, so that we may take $B_{i,m} = A_i$ in assumption A-1).

Natural candidates for the covering sets $C_{i,m}$ are $\sqrt{2\epsilon(m)}$ dilations of the A_i. That is, let $\delta_m = \sqrt{2\epsilon(m)}$ and define

$$\bar{C}_{i,m} = \{\nu \in \mathcal{M}_1(\Sigma^{i+m}) : \ d(\nu, (\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i) < \delta_m \text{ for some Markov kernel } \pi^i\}$$

and let

$$C_{i,m} = \{\nu \in \mathcal{M}_1(\Omega) : \ \nu_{i+m} \in \bar{C}_{i,m}\}\,.$$

It is clear that $C_{i,m}$ is open, and also that

$$\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} (C_{i,m}^{\sqrt{2\epsilon(m)}} \backslash A_i) = \emptyset.$$

Therefore, by using Theorem 4 and the procedure described in Theorem 3, the sets $C_{i,m}$ are candidates for building a decision rule which, a.s., decides correctly in finite time whether the given observation sequence was generated by a Markov chain of order i. In order to be able to do so, we need only to check that the complements of the sets $C_{i,m}$, which are closed, have the property that they may be covered by small enough spheres (say, $\delta_m/4$ spheres), such that the union of those spheres belongs to the complement of some $C_{i,m'}$. This can be seen by using the following lemma.

Lemma 2 Let $\nu, \nu' \in \mathcal{M}_1(\Sigma^k)$. Assume that for some π^i, $i \leq k-2$,

$$d(\nu, (\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i) > \delta_m.$$

Further assume that $d(\nu, \nu') < \delta_m/4$. Then

$$d(\nu', (\nu')_i \otimes \pi^i \otimes \cdots \otimes \pi^i) > \delta_m/2.$$

Proof Note that $d(\nu, \nu') < \delta_m/4$ implies that $d((\nu)_i, (\nu')_i) < \delta_m/4$ and that $d((\nu)_{i+1}, (\nu')_{i+1}) < \delta_m/4$. On the other hand, since π^i is a Markov kernel, it also follows that $d((\nu)_i \otimes \pi^i, (\nu')_i \otimes \pi^i) < \delta_m/4$ and therefore also that $d((\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i, (\nu')_i \otimes \pi^i \otimes \cdots \otimes \pi^i) < \delta_m/4$. Hence,

$$\begin{aligned} d(\nu', (\nu')_i \otimes \pi^i \otimes \cdots \otimes \pi^i) &\geq d(\nu, (\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i) - d(\nu, \nu') \\ &\quad - d((\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i, (\nu')_i \otimes \pi^i \otimes \cdots \otimes \pi^i) \\ &> \delta_m/2 \end{aligned}$$

□

We have now completed all the preparatory steps required for the definition of the proposed decision rule. Indeed, let $\epsilon_i(m)$ be a sequence of positive numbers, define

$$\alpha_i(m) = \frac{2^{4m+7}}{\epsilon_i(m)} \left[3 \log m + N^{\Sigma^{i+m}} \left(\sqrt{\epsilon_i(m)/2^{4m+3}} \right) \left(1 - \log \sqrt{\epsilon_i(m)/2^{4m+3}} \right) \right]\,. \tag{4.35}$$

We have, by Lemma 2 and Theorem 4, that for any Markov measure $\mu \in A_i$,

$$P_\mu(\mu^{i+m}_{\alpha_i(m)} \in C^c_{i,m}) \leq \frac{1}{m^2} \tag{4.36}$$

where we have used (4.27). The construction of the decision rule is then identical to the one described in Theorem 3, i.e. one forms the parsing of the observation sequence into the nonoverlapping blocks X^m described in equation (3.20) with $\beta(m)$ chosen as in (3.21). At each step, one forms, based on the block X^m, the empirical measures of order $m, m+1, \ldots, 2m$. The order estimate at the m-th step is now the smallest i such that $\mu^{i+m}_{\alpha_i(m)} \in C_{i,m}$. By the results of section 3, this decision rule achieves a.s. only a finite number of errors, regardless of the true order.

Appendix

For completeness, here we prove a lower bound for the covering number of $\mathcal{M}_1(\Sigma)$ with respect to the Prohorov metric. This lower bound exhibits a behavior similar to the upper bound of (2.8), so that these bounds cannot be much improved. In the proof below, $M(\epsilon, Y, \eta)$ denotes the ϵ-capacity (or packing number) of the space Y with respect to the metric η. That is, $M(\epsilon, Y, \eta)$ represents the maximum number of non-overlapping balls of *diameter* ϵ with respect to the metric η that can be packed in Y. The well known relationship

$$N(2\epsilon, Y, \eta) \leq M(2\epsilon, Y, \eta) \leq N(\epsilon, Y, \eta)$$

between covering numbers and packing numbers is easy to show and is used in the proof below. Note that for a Polish space Σ with metric η, we use the notations $N(\epsilon, \Sigma, \eta) = N^{\Sigma}(\epsilon)$ and $N(\epsilon, \mathcal{M}^1(\Sigma), d) = N(\epsilon, \mathcal{M}^1(\Sigma))$.

Lemma: Let Σ be compact Polish space with metric η, and let $\mathcal{M}^1(\Sigma)$ denote the set of probability measures on Σ with the Prohorov metric d. Then

$$N(\epsilon, \mathcal{M}^1(\Sigma)) \geq 8\epsilon\sqrt{N^{\Sigma}(2\epsilon)}\left(\frac{1}{8\epsilon}\right)^{N^{\Sigma}(2\epsilon)}$$

Proof: First, we can find $N = N^{\Sigma}(\epsilon)$ points $x_1, \ldots, x_N$ which are pairwise greater than or equal to ϵ apart. Each measure supported on these N points corresponds to a point in $\mathbb{R}^N$ in the natural way. Then, the set of all probability measures supported on $x_1, \ldots, x_N$ corresponds to the simplex S^N in $\mathbb{R}^N$.

Now, let p, q be points on the simplex S^N and suppose that $d_{\ell^1}(p,q) \geq 2\epsilon$ where $d_{\ell^1} = \sum_{i=1}^N |p_i - q_i|$. Then on some subset $G \subset \{1, \ldots, N\}$ of coordinates either $\sum_{i \in G} p_i \leq \sum_{i \in G} q_i + \epsilon$ or $\sum_{i \in G} q_i \leq \sum_{i \in G} p_i + \epsilon$. Then, considered as probability measures on Σ, $d(p,q) \geq \epsilon$ since there is a closed set $F \subset \Sigma$, namely $F = \{x_i \mid i \in G\}$, for which either $p(F) \geq q(F^{\epsilon}) + \epsilon$ or $q(F) \geq p(F^{\epsilon}) + \epsilon$. Hence,

$$N(\epsilon/2, \mathcal{M}^1(\Sigma), d) \geq M(\epsilon, \mathcal{M}^1(\Sigma), d) \geq M(2\epsilon, S^N, d_{\ell^1}) \geq N(2\epsilon, S^N, d_{\ell^1})$$

Finally, to get a lower bound on $N(2\epsilon, S^N, d_{\ell^1})$, we note that the $N-1$ dimensional surface measure of the simplex S^N is $\sqrt{N}/(N-1)!$ (simply, differentiate the N-dimensional volume of the interior of an

x-scaled simplex with respect to x, taking the angles into account). On the other hand, note that the $N-1$ dimensional volume of the intersection of S^N with an N dimensional ℓ^1 ball of radius 2ϵ is not larger than the volume of an $N-1$ dimensional ℓ^1 ball of radius 2ϵ, which equals $(4\epsilon)^{N-1}/(N-1)!$. Thus, $N(2\epsilon, S^N, d_{\ell^1}) \geq (1/4\epsilon)^{N-1}\sqrt{N}$. Thus, $N(\epsilon/2, \mathcal{M}^1(\Sigma)) \geq (1/4\epsilon)^{N^{\Sigma}(\epsilon)} 4\epsilon\sqrt{N^{\Sigma}(\epsilon)}$, or equivalently $N(\epsilon, \mathcal{M}^1(\Sigma)) \geq 8\epsilon\sqrt{N^{\Sigma}(2\epsilon)}(1/8\epsilon)^{N^{\Sigma}(2\epsilon)}$.

□

References

[1] Barron, A.R. (1985). Logically Smooth Density Estimation. Ph.D. thesis, Dept. of Electrical Engineering, Stanford University, Sept.

[2] Barron, A.R. and Cover, T.M. (1991). Minimum complexity density estimation. *IEEE Trans. Info. Theory*, Vol. 37, pp. 1034-1054.

[3] Cover, T.M. (1972). A hierarchy of probability density function estimates. In *Frontiers of Pattern Recognition*, Academic Press Inc.

[4] Cover, T.M. (1973). On determining the irrationality of the mean of a random variable. *The Annals of Statistics*, Vol 1, pp. 862-871.

[5] Dembo, A. and Peres, Y. (1991). A topological criterion for hypothesis testing. To appear, *The Annals of Statistics.*

[6] Deuschel, J.D. and Stroock, D.W. (1989). *Large Deviations.* Academic Press, Boston.

[7] Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression. *J. Roy. Stat. Soc.*, Ser. B., 41, pp. 190-195.

[8] Hannan, E.J. (1980). The estimation of the order of an ARMA process. *The Annals of Statistics*, Vol. 8, pp. 1071-1081.

[9] Ibragimov, I.A. and Has'minskii, R.Z. (1981). *Statistical Estimation*, Springer.

[10] Koplowitz, J., (1977). Abstracts of papers, *Int. Symp. on Info. Theory*, Cornell Univ., Ithaca, NY, Oct. 10-14, p. 64.

[11] Kulkarni, S.R. and Zeitouni, O. (1991). Can one decide the type of the mean from the empirical measure? *Stat. & Prob. Letters*, Vol. 12, pp. 323-327, 1991.

[12] Merhav, N., Gutman, M. and Ziv, J. (1989). On the determination of the order of a Markov chain and universal data compression. *IEEE Trans. Info. Theory*, Vol. IT-35, pp. 1014-1019.

[13] Shibata, R.. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, Vol 8, pp. 147-164.

[14] Sion, M. (1958). On general minimax theorems. *Pacific J. Math.*, Vol. 8, pp. 171-175.